ANALYSIS AND IMPROVEMENT OF THE VECTOR QUANTIZATION IN SELP

W. B. KLEIJN, D. J. KRASINSKI, and R. H. KETCHUM

AT&T BELL LABORATORIES
200 Park Plaza
Naperville, IL 60510, USA

ABSTRACT

The SELP algorithm is described as a speech coding method employing a two-stage vector quantization. The first stage uses an adaptive codebook which efficiently encodes the periodicity of voiced speech, and the second stage uses a stochastic codebook to encode the remainder of the excitation signal. The adaptive codebook performs well when the pitch period of the speech signal is larger than the frame size. An extension is introduced, which increases its performance for the case that the frame size is longer than the pitch period. The performance of the stochastic stage, which improves with frame length, is shown to be best in those sections of the speech signal where a high level of short-term correlation is present. It can be concluded that the SELP algorithm performs best during voiced speech where the pitch period is longer than the frame length.

1. INTRODUCTION

The Stochastically Excited Linear Prediction (SELP) algorithm for speech coding, originally proposed by Atal and Schroeder [1], is a method of vector quantizing the sampled speech signal on a frame-by-frame basis. The algorithm vector quantizes a signal which, ideally, contains no short-term correlation. When the quantized signal is filtered through an appropriate linear filter (determined with LPC methods) to provide the correct short-term correlation, synthetic speech results. The signal can be synthesized at the receive end of the speech coder by transmitting both an index to the particular excitation vector selected from the codebook and a description of the filter.

The effect of the present frame excitation on the synthetic speech in the present frame can be expressed as a convolution of the excitation with the filter impulse response. This convolution can be implemented as a matrix multiplication of the excitation vector with a lower-triangular Toeplitz matrix $\mathbf{H}$ containing the filter impulse response along the first column. The synthetic speech signal $\mathbf{y}$ produced by a candidate excitation vector $\mathbf{r}$ is now:

$$\mathbf{y} = \mu \mathbf{H} \mathbf{r} + \mathbf{z}, \tag{1}$$

where μ is a scaling factor, and $\mathbf{z}$ represents the zero-input response of the filter describing the present frame response to the preceding excitation.

Ideally the excitation vector $\mathbf{r}$ results in a synthetic speech signal identical to the original speech. It is this ideal "target" excitation vector, $\mathbf{t}$, which is vector quantized in the SELP algorithm. The squared error between the original speech signal and the synthetic speech signal can be used as the quantization error criterion. Expressed in terms of excitation vectors, it becomes:

$$\epsilon = (\mathbf{t}-\mu\mathbf{r})^T\mathbf{H}^T\mathbf{H}(\mathbf{t}-\mu\mathbf{r}). \quad (2)$$

The error criterion described represents the squared error of the synthetic speech signal. It does not consider that uniform matching of the speech signal over its entire spectrum does not minimize the perceived noise level. The error signal is less perceptible in those sections of the spectrum which contain relatively high energy (the formants). Thus, it is advantageous to move more of the error signal under the formants, while reducing the error signal in other regions [2]. This can be accomplished by moving the poles of the filter inward by a factor γ, which is chosen in the range between 0.7 and 0.9. An added advantage of this "noise-weighting" procedure is that it dampens the impulse response significantly, so that it can be truncated at a relatively short length. From here on we assume that the matrix **H** includes noise weighting.

The development of fast algorithms benefits from increased symmetry. By truncating the impulse response after **R** samples, and considering this entire response length for each sample of the excitation vector in the error criterion, a more symmetric spectral-weighting matrix $\mathbf{H}^T\mathbf{H}$ can be created [3]. This modified error criterion compares the response of the excitation vector to that of the target excitation vector in the present as well as subsequent frames. The resulting $\mathbf{H}^T\mathbf{H}$ matrix is still **NxN** but is now a Toeplitz band matrix. A fast algorithm which takes advantage of this symmetry is described in [3].

Two codebooks are used in our SELP implementation. The first codebook is adaptive, adjusting to the periodicity of the speech signal. The method is similar to procedures described in [4] and [5]. The adaptive codebook is constructed by buffering the synthetic excitation. If **M** samples of past excitation are stored, and the frame length is **N** samples, then the adaptive codebook contains **M−N+1** distinct candidate vectors. Neighboring candidate vectors differ by a shift of one sample.

Because of the linearity of the filtering procedure, the excitation vector selected from the adaptive codebook can be subtracted from the target excitation vector **t** to obtain a new target vector for the second vector quantization procedure. Since the LPC filter models the short-term correlation, and the adaptive codebook the long-term correlation of the speech signal, the new target vector has a noisy character. The stochastic codebook consists of a set of vectors containing normally distributed samples. It is often center clipped to increase computational speed [3], [6], [7].

During the vector quantization procedures of both the adaptive and stochastic codebooks all candidate vectors are scaled by optimal scaling factors. This allows for the large dynamic range present in speech. It is easily shown from equation (2) that the optimal scaling factor μ for a candidate vector **r** takes the value $\mathbf{r}^T\mathbf{H}^T\mathbf{H}\mathbf{t}/\mathbf{r}\ ^T\mathbf{H}^T\mathbf{H}\mathbf{r}$. The scaling factors also must be transmitted to the synthesizer.

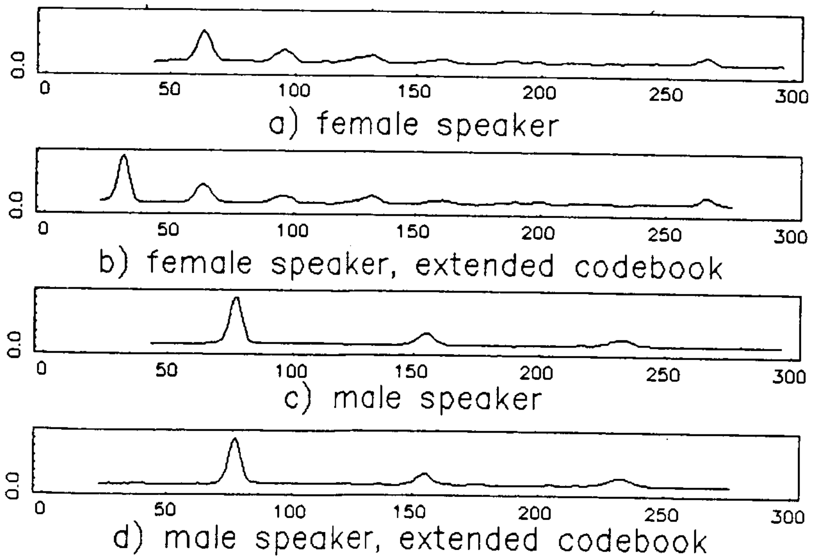
2. ANALYSIS AND IMPROVEMENT OF SELP

2.1 The Adaptive Codebook

The adaptive codebook procedure performs best in voiced sections of speech. The codebook vector selected usually contains a section of synthetic excitation that had been applied one or more integer number of pitch periods prior to the present frame. Figs. 1a and 1c illustrate the probability density function for the delay (in samples) between the selected candidate vector and the present vector for speech from a female and a male speaker, respectively. The probability density function for each speaker was approximated from an analysis of 10000 frames of 8 kHz sampled speech using a SELP coder with two codebooks of 256 vectors and a frame length of 40 samples. The pitch period of the speakers is clearly displayed in these distributions. For the male speaker a delay of one pitch period is most prevalent, but two and three pitch delays contribute significantly also. For the female speaker, a delay of one pitch period cannot be chosen, because the frame length is longer than the pitch period. Instead, delays of two through five integer pitch periods are prevalent.

The subjective performance of the SELP coder is better for the speech signal with a pitch period longer than the frame length. However, we have found that, at fixed codebook size, the segmental signal to noise ratio decreases smoothly as a function of increasing frame length. Thus, the modeling accuracy is not affected significantly by the frame length being longer or shorter than the pitch period. However, only if the frame length is shorter than the pitch period will noise resulting from inaccurate modeling be repeated at the same rate as the pitch pulses. This leads to desirable harmonics in spectral regions which are numerically unimportant in the evaluation of the error criterion. The harmonics at the high end of the spectrum, which contains less energy, are often created from noise signals. This effect is eliminated when the pitch period is shorter than the frame length.

Since (at constant codebook size) the bit rate of the SELP coder is inversely related to frame length, it is desirable to minimize the negative impact of increased frame length. This can be accomplished with an extension of the adaptive codebook which results in enhanced periodicity [3]. In the conventional method, the candidate vector containing the most recent synthetic excitation starts exactly N samples before the present frame. A candidate vector starting at a more recent sample is incomplete. However, if the existing portion of such a vector is presented in the first part of the present frame, it will create a periodicity. If the vector starts p samples previous to the present frame, then the periodicity will be p samples. This periodicity can be maintained by repeating the existing portion of the vector again, starting p samples into the present frame. Thus, the adaptive codebook can be extended with "virtual" vectors with a period equal to the length of the existing portion.



a) female speaker

b) female speaker, extended codebook

c) male speaker

d) male speaker, extended codebook

Fig. 1. Probability Density of Delay

At an equal number of candidate entries, the extended adaptive codebook results in an improved subjective performance when the frame length is longer than a single pitch period. Fig. 1b displays the probability density function for the delay of the selected vectors for the extended codebook case. The candidate vectors chosen most often are those with a delay of one pitch period, which were completed in the manner described above. This means that in a large number of frames, the newly created periodic entries in the codebook provided the best match to the target excitation vector **t**. In Fig. 2 spectra generated with and without the extension are shown. The virtual candidates create a periodicity, resulting in enhanced harmonicity of the spectrum, especially between 1200 and 3000 Hz. It is this enhanced periodicity which improves the subjective speech quality. However, the associated segmental signal to noise ratio increased only about 0.1 dB. The improvement of the subjective speech quality for the female speaker is in contrast to the results for the male speaker (Fig. 1d), for whom the revised codebook with the modified range of available delays has no noticeable effect on overall performance.

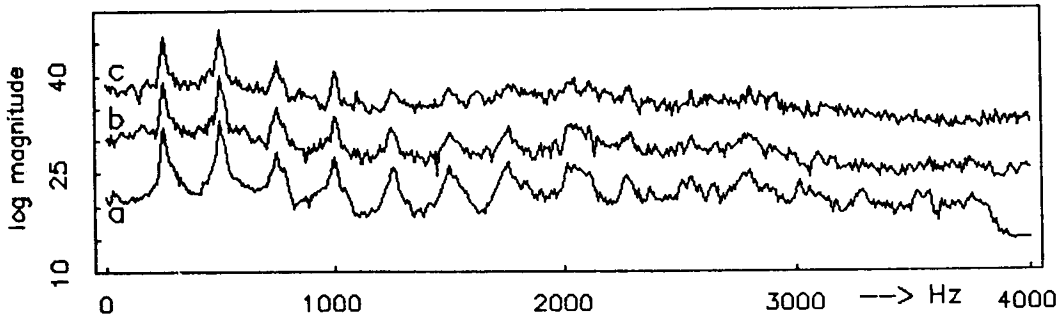


Fig. 2. Spectral Characteristics
a: original, b: extended codebook, c: conventional codebook

2.2 The Stochastic Codebook.

The performance of the stochastic codebook can be evaluated by assuming that the part of the excitation signal which remains after the first vector quantization stage contains normally distributed random samples. Its error criterion is simplified when expressed in the eigenspace of the $\mathbf{H}^T\mathbf{H}$ matrix. The diagonal transform of the spectral weighting matrix is $\Lambda=\mathbf{U}\mathbf{H}^T\mathbf{H}\mathbf{U}^T$, where **U** is a unitary matrix. The transforms of the excitation vectors are $\tau=\mathbf{U}\mathbf{t}$ and $\rho=\mathbf{U}\mathbf{r}$ respectively. Since the vectors **t** and **r** were assumed to be sequences of normally distributed independent samples, and since $\mathbf{H}^T\mathbf{H}$ is positive definite, τ and ρ are also normally distributed independent sequences. Equation (2) can now be written as:

$$\epsilon = (\tau-\mu\rho)^T\Lambda(\tau-\mu\rho). \qquad (3)$$

The behavior of the error criterion ϵ is a function of the framelength **N**, the eigenvalues described by Λ, and codebook size.

When trying to match noise signals, the matrix $\mathbf{H}^T\mathbf{H}$ tends to be close to the identity matrix. Fig. 3 shows the dependence of the minimized error criterion ϵ, normalized to a per sample basis, on the frame length for a weighting matrix whose eigenvalues are all equal. Six error curves are shown for codebooks of various sizes. In addition, the error as a function of frame size under the constraint of constant bit rate for the codebook index is shown. From the latter curves it appears that the coding efficiency is identical at all frame lengths. This is not the case, however, since additional bits must be spent on the encoding of the scaling factor used in the quantization procedure. Since this overhead becomes smaller with increasing frame length, the encoding becomes more efficient at larger frame lengths.

A high level of short-term correlation results in a large range for the magnitudes of the eigenvalues of the spectral weighting matrix. Then, only a small fraction of the components of the target vector r (those associated with large eigenvalues) are of importance in the evaluation of the error criterion. It is illuminating to consider the extreme instance where only one eigenvalue is nonzero. In this case only a single component of the target vector r must be matched. Using an optimal scaling factor, any vector not orthogonal to this component will result in zero error. However, the quantization of this scaling factor will introduce error in an actual coder.

Using these two extreme cases as guideline, it is seen that the stochastic quantization procedure can be expected to perform best in voiced speech, which has pronounced formants. Thus, the stochastic codebook search will perform best in the same voiced regions where the adaptive codebook search performs best. These results are consistent with intuition; the speech coders perform best where the signal has the most structure (low entropy, low information content), and worst where the signal has the least structure (high entropy, high information content). These results lead to segmental signal to noise ratios which are high during speech and low during background noise.

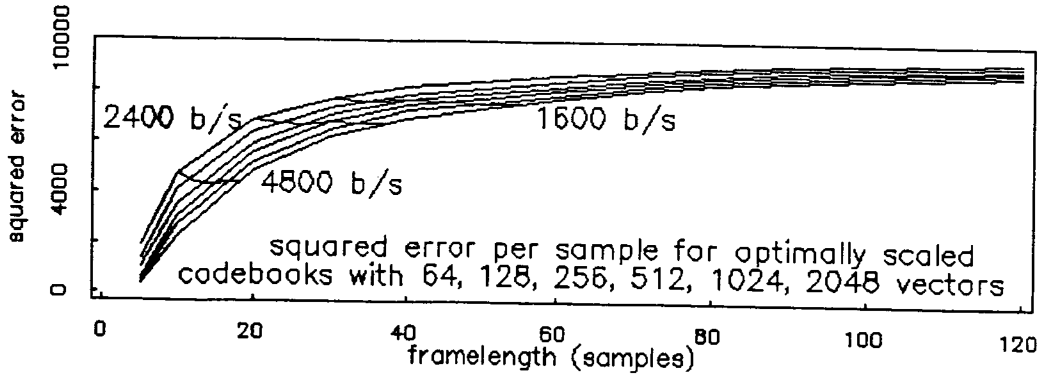


Fig. 3. Squared Error for Noise Signal

3. CONCLUSION

The encoding of the periodicity of the excitation signal of SELP coders deteriorates when the frame length exceeds the length of a pitch period. This effect can be diminished by extending the adaptive codebook with vectors of progressively shorter periodicity. The performance of the stochastic codebook improves when the frame length is increased. Both the adaptive and stochastic codebook quantization procedure perform best during voiced speech where both short-term and long-term correlation are highest.

It is natural that the adaptive codebook quantization is performed before the stochastic quantization, with the latter acting as a correction to an already good excitation sequence. Because of the recursive character of the adaptive codebook, it is logical to apply the stochastic codebook correction before entering the excitation into the adaptive codebook.

This means that the stochastic codebook frame length must fit an integer number of times in the frame length used for the adaptive codebook. This and the fact that the stochastic codebook performs better with increasing length means that the frame length of the stochastic codebook is best chosen to be equal to that of the adaptive codebook.

REFERENCES

[1] Atal B.S. and M.R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates", *Proc. of ICC*, Amsterdam, 1610-1613, 1984.

[2] Atal, B.S. and M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria" *IEEE Trans. Speech Signal Proc. Vol. ASSP-27, no 3*, 247-254, 1979.

[3] Kleijn, W.B., D.J. Krasinski, and R.H. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP" *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, New York, 1988.

[4] Singhal S. and B.S. Atal, "Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, San Diego, 1.3.1-1.3.4, 1984.

[5] Rose, R.C. and T.P. Barnwell, "Quality Comparison of Low Complexity 4800bps Self Excited and Code Excited Vocoders", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Dallas, 1637-1640, 1987.

[6] Lin, D., "New approaches to Stochastic Coding of Speech Sources at Very Low Bit Rates", in *Signal Processing III: Theories and Applications*, I.T. Young et al. eds., Elsevier, 445-447, 1986.

[7] Davidson, G. and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Tokyo, 3055-3058, 1986.